Homework 4 - Samuel Fafel

Problem 1: Naïve Bayes for Text Classificaiton (15 points)

1) Consider the documents with word counts $x = (1, 0, 1)$, i.e., count for w_1 is 1, count for w_2 is 0, count for w_3 is 1. Which class has the highest posterior probability?

2) Consider the documents with word counts $x = (2, 0, 1)$, i.e., count for w_1 is 2, count for w_2 is 0, count for w_3 is 1. Which class has the highest posterior probability?

1) $x = (1, 0, 1)$

For class $y = $ -1:

$$\Pr(x|y = \text{-}1) = \Pr(w_1|y = \text{-}1) \cdot \Pr(w_3|y = \text{-}1) = \frac{1}{10} \cdot \frac{7}{10}$$

$$\Pr(y = \text{-}1 \mid x) = \frac{\frac{1}{10} \cdot \frac{7}{10} \cdot \frac{4}{10}}{\Pr(x)} = \frac{\frac{28}{1000}}{\Pr(x)}$$

For class $y = 1$:

$$\Pr(y = 1|x) = \Pr(w_1|y = 1) \cdot \Pr(w_3|y = 1) = \frac{5}{10} \cdot \frac{3}{10}$$

$$\Pr(y = 1 \mid x) = \frac{\frac{5}{10} \cdot \frac{3}{10} \cdot \frac{6}{10}}{\Pr(x)} = \frac{\frac{90}{1000}}{\Pr(x)}$$

Since $\Pr(x)$ is the same for both classes, we can just compare the numerators of Bayes' Theorem:

For class $y = $ -1: $\frac{28}{1000}$

For class $y = 1$: $\frac{90}{1000}$

Therefore the class with the highest posterior probability for $x = (1, 0, 1)$ is $y = 1$.

2) $x = (2, 0, 1)$

For class $y = $ -1:

$$\Pr(x \mid y = \text{-}1) = \Pr(w_1|y = \text{-}1)^2 \cdot \Pr(w_3|y = \text{-}1) = \left(\frac{1}{10}\right)^2 \cdot \frac{7}{10}$$

$$\Pr(y = \text{-}1 \mid x) = \frac{\left(\frac{1}{10}\right)^2 \cdot \frac{7}{10} \cdot \frac{4}{10}}{\Pr(x)} = \frac{\frac{28}{10000}}{\Pr(x)}$$

For class $y = 1$:

$$\Pr(y = 1 \mid x) = \Pr(w_1|y = 1)^2 \cdot \Pr(w_3|y = 1) = \left(\frac{5}{10}\right)^2 \cdot \frac{3}{10}$$

$$\Pr(y = 1 \mid x) = \frac{\left(\frac{5}{10}\right)^2 \cdot \frac{3}{10} \cdot \frac{6}{10}}{\Pr(x)} = \frac{\frac{450}{10000}}{\Pr(x)}$$

Since $\Pr(x)$ is the same for both classes, we can just compare the numerators of Bayes' Theorem:

For class $y = $ -1: $\frac{28}{10000}$

For class $y = 1$: $\frac{450}{10000}$

Therefore the class with the highest posterior probability for $x = (2, 0, 1)$ is $y = 1$.

Homework 4 - Samuel Fafel

Problem 2: Implementing Naïve Bayes for Text Classification (50 points)

Requirements 1) The current Naïve Bayes classifier was built up without using Laplacian smoothing, i.e., it currently uses equation (1). You need to locate the right place in the code to incorporate Laplacian smoothing as equation (2).

2) Apply the learned classifier to predict the class labels for the test documents. Report the classification accuracy over the test documents (i.e., the proportion of test documents that are classified correctly).

See p2.py

Reported Classification Accuracy: 0.7811 (78.11%)

Problem 3: Regularized Logistic Regression (35 points)

You are required to:

- Use the 5-fold cross validation method to decide the best value of the parameter C. The candidate values for C are $0.01, 0.1, 1, 10, 100, 1000$. For each C, report the training error and validation error. Choose the best C that yields the lowest validation error.
- Use the selected best C value to train a logistic regression model on the whole training data and evaluate and report its performance (by accuracy) on the testing data.
- Report the results.

Values for C: [0.01, 0.1, 1, 10, 100, 1000]
Best C: 10
Validation Errors: [0.426, 0.4, 0.393, 0.38, 0.393, 0.387]
Training Errors: [0.218, 0.137, 0.0583, 0.017, 0.0, 0.0]
Final Accuracy: 0.7758620689655172 (77.58%)